DOCUMENT RESUME

ED 268 138 TM 850 775

| | |
|---|---|
| AUTHOR | Mislevy, Robert J. |
| TITLE | Bayes Modal Estimation in Item Response Models. |
| INSTITUTION | Educational Testing Service, Princeton, N.J. |
| SPONS AGENCY | Spencer Foundation, Chicago, Ill. |
| REPORT NO | ETS-RR-85-33 |
| PUB DATE | Aug 85 |
| NOTE | 57p. |
| PUB TYPE | Reports - Research/Technical (143) |
| EDRS PRICE | MF01/PC03 Plus Postage. |
| DESCRIPTORS | Algorithms; *Bayesian Statistics; *Estimation (Mathematics); *Latent Trait Theory; *Mathematical Models; Maximum Likelihood Statistics; *Psychometrics; Simulation; Statistical Analysis; Statistical Data |
| IDENTIFIERS | *EM Algorithm; *Item Parameters; One Parameter Model; Population Parameters; Three Parameter Model; Two Parameter Model |

ABSTRACT

Simultaneous estimation of many parameters can often be improved, sometimes dramatically so, if it is reasonable to consider one or more subsets of parameters as exchangeable members of corresponding populations. While each observation may provide limited information about the parameters it is modeled directly in terms of, it also contributes information about the populations to which they belong. Knowledge about the populations, generally superior to knowledge about individual parameters, can in turn be brought to bear in the estimation of any individual parameter. This article describes a Bayesian framework for estimation in item response models, with two-stage prior distributions on both item and examinee populations. Strategies for point and interval estimation are discussed and a general procedure based on the EM algorithm is presented. Details are given for implementation under one-, two-, and three-parameter logistic item response theory models. Novel features include minimally restrictive assumptions about examinee distributions and the exploitation of dependence among item parameters in a population of interest. Improved estimation in a moderately small sample is demonstrated with simulated data. Possible extensions of the procedures are discussed. (Author/PN)

**********
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
**********

RR-85-33

ED268138

RESEARCH REPORT

# BAYES MODAL ESTIMATION IN ITEM RESPONSE MODELS

Robert J. Mislevy

U.S. DEPARTMENT OF EDUCATION
NATIONAL INSTITUTE OF EDUCATION
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

☒ This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official NIE position or policy.

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

P. Feldmesser

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

TM 850 475

Educational Testing Service
Princeton, New Jersey
August 1985

# Bayes Modal Estimation in Item Response Models*

Robert J. Mislevy

Educational Testing Service

Princeton, NJ

August 1985

*This research was supported by a grant from the Spencer Foundation, Chicago, IL.

Copyright © 1985. Educational Testing Service. All rights reserved.

Abstract

This article describes a Bayesian framework for estimation in item response models, with two-stage prior distributions on both item and examinee populations. Strategies for point and interval estimation are discussed, and a general procedure based on the EM algorithm is presented. Details are given for implementation under one-, two-, and three-parameter logistic IRT models. Novel features include minimally restrictive assumptions about examinee distributions and the exploitation of dependence among item parameters in a population of interest. Improved estimation in a moderately small sample is demonstrated with simulated data.

Key words: Bayesian estimation
EM algorithm
Hierarchial prior distributions
Item response models
Marginal maximum likelihood

## Introduction

Simultaneous estimation of many parameters can often be improved, sometimes dramatically so, if it is reasonable to consider one or more subsets of parameters as exchangeable members of corresponding populations (Efron & Morris, 1975; James & Stein, 1961; Kelley, 1927; Lindley & Smith, 1972). The idea is that while each observation may provide limited information about the parameters it is modeled directly in terms of, it also contributes information about the populations to which they belong. Knowledge about the populations, generally superior to knowledge about individual parameters, can in turn be brought to bear in the estimation of any individual parameter. Novick et al. (1972) and Rubin (1980), for example, provide Bayes and empirical Bayes solutions respectively to the problem of predicting student performance in a given law school when data are available for several law schools. Both studies obtained more stable estimates in small schools and improved cross-validation results when compared to independent estimation within schools.

Analogous procedures for the IRT setting have begun to appear in the psychometric literature. Bock and Aitkin (1981), Rigdon and Tsutakawa (1983), and Thissen (1982) address the problem of incidental examinee parameters by integrating over a population density to produce marginal likelihood functions for item

parameters. Reiser (1981) and Mislevy and Bock (1981) extended this model by positing prior distributions for item parameters. Swaminathan and Gifford (1982, 1984, in press) employ two-stage priors for examinee parameters and selected item parameters, then obtain the joint posterior mode for all individual parameters. Andersen and Madsen (1977), Mislevy (1984), and Sanathanan and Blumenthal (1978) provide maximum likelihood solutions for the parameters of examinee population distributions, conditional on item parameters. Finally, Bock and Aitkin (1981) and Bock and Mislevy (1982) derive posterior means and standard deviations of the parameters of individual examinees, conditional on item and examinee population parameters.

The aformentioned procedures can all be expressed as special cases of a more comprehensive Bayesian framework for estimation in item response models. Working along lines first suggested by Lindley and Smith (1972), we begin by introducing a model for item responses that employs two levels of prior distributions on both item and examinee parameters. A general discussion of theoretical and practical considerations in estimating the parameters of such a model, including an EM computing algorithm (Dempster, Laird, & Rubin, 1977), follows. Procedures specific to logistic item response models (Birnbaum, 1968; Rasch, 1960a; Lord, 1980) are then detailed. We illustrate the techniques with simulated data and conclude by discussing possible extensions of the procedures.

## The General Form of the Model

Let $\theta$ denote examinee ability and $p(\theta|\underset{\sim}{\tau})$ its density, conditional on examinee population parameters $\underset{\sim}{\tau}$. If $\theta$ follows a normal distribution, for example, $\underset{\sim}{\tau} = (\mu_\theta, \sigma^2_\theta)$, the mean and variance. $\underset{\sim}{\tau}$ is assumed in turn to follow density $p(\underset{\sim}{\tau})$. In the same manner, let $\xi$ denote the parameter(s) of a test item and $p(\xi|\underset{\sim}{\eta})$ denote its density, conditional on item population parameters $\underset{\sim}{\eta}$; $\underset{\sim}{\eta}$ in turn follows density $p(\underset{\sim}{\eta})$. Independence over examinees and items is assumed, given $\underset{\sim}{\tau}$ and $\underset{\sim}{\eta}$.

Let $d_{ij}$ take the value 1 if examinee i is administered item j and 0 if not. For n items of interest, let $\underset{\sim}{d}_i = (d_{i1},\ldots,d_{in})$, and for N examinees, let $\underset{\sim}{D} = (\underset{\sim}{d}_1,\ldots,\underset{\sim}{d}_N)$. Let $u_{ij}$ denote the response of examinee i to item j, taking the value 1 if the item was adminstered and answered correctly, and 0 otherwise; define $\underset{\sim}{u}_i$ and $\underset{\sim}{U}$ in analogy to $\underset{\sim}{d}_i$ and $\underset{\sim}{D}$. Denote by $L(\underset{\sim}{U}|\underset{\sim}{D},\underset{\sim}{\theta},\underset{\sim}{\xi})$ the likelihood of of the possibly incomplete matrix of responses of subjects with abilities $\underset{\sim}{\theta} = (\theta_1,\ldots,\theta_N)$ to items with parameters $\underset{\sim}{\xi} = (\xi_1,\ldots,\xi_n)$. By Bayes theorem, the posterior density of $\underset{\sim}{\theta}$, $\underset{\sim}{\xi}$, $\underset{\sim}{\tau}$, and $\underset{\sim}{\eta}$, given realized observations $\underset{\sim}{U}$ is given by

$$p(\underset{\sim}{\theta},\underset{\sim}{\xi},\underset{\sim}{\tau},\underset{\sim}{\eta}|\underset{\sim}{D},\underset{\sim}{U}) \propto L(\underset{\sim}{U}|\underset{\sim}{D},\underset{\sim}{\theta},\underset{\sim}{\xi}) \cdot p(\underset{\sim}{\theta}|\underset{\sim}{\tau}) \cdot p(\underset{\sim}{\tau}) \cdot p(\underset{\sim}{\xi}|\underset{\sim}{\eta}) \cdot p(\underset{\sim}{\eta}) \quad . \tag{2.1}$$

After the forms of the likelihood function L and the prior densities $p(\underset{\sim}{\theta}|\underset{\sim}{\tau})$ and $p(\underset{\sim}{\xi}|\underset{\sim}{\eta})$ have been chosen, the highest level prior densities $p(\underset{\sim}{\tau})$ and $p(\underset{\sim}{\eta})$ have been specified, and the data $\underset{\sim}{U}$ have been observed, (2.1) contains all information available about the parameters in the model. The sheer incomprehensibility of a joint distribution of possibly thousands of variables, however, demands summary in terms of salient attributes, to be used in constructing point and interval estimates, for example.

The joint mean of the posterior has the desirable property that the value for each component retains the same value in any marginal distribution obtained by integrating (2.1) over any subset of remaining components. Posterior modes, which do not exhibit this invariance, are more often seen in practice in complex problems such as the one at hand, since they prove easier to obtain. Generally speaking, a parameter's marginal posterior mode is a better approximation of its posterior mean than is its joint mode (O'Hagan, 1976). This is especially so when "nuisance" parameters appearing in the joint posterior, along with the parameters of interest, are poorly determined. Examinee parameters $\underset{\sim}{\theta}$ follow this description in the present context, and we shall integrate over their distribution routinely to obtain

$$p(\underset{\sim}{\xi},\underset{\sim}{\tau},\underset{\sim}{\eta}|\underset{\sim}{D},\underset{\sim}{U}) = \int_{\underset{\sim}{\theta}} p(\underset{\sim}{\theta},\underset{\sim}{\xi},\underset{\sim}{\tau},\underset{\sim}{\eta}|\underset{\sim}{D},\underset{\sim}{U})\, d\underset{\sim}{\theta} \quad . \tag{2.2}$$

The reduction in dimensionality thus achieved assures that the marginal modes of the remaining item and population parameters will better approximate their means.

In principle, it is also possible to integrate over item parameters as well in order to obtain the marginal distributions of item and examinee population parameters alone:

$$p(\underset{\sim}{\tau},\underset{\sim}{\eta}|\underset{\sim}{D},\underset{\sim}{U}) \propto \int_{\underset{\sim}{\xi}} \int_{\underset{\sim}{\theta}} p(\underset{\sim}{\theta},\underset{\sim}{\xi},\underset{\sim}{\tau},\underset{\sim}{\eta}|\underset{\sim}{D},\underset{\sim}{U})\, d\underset{\sim}{\theta}\, d\underset{\sim}{\xi} \quad . \tag{2.3}$$

The numerical integration required to effect (2.3), however, is not tractible for any but trivial problems with currently available computing machinery. An alternative suggested by Leonard (1982) is to approximate the marginal density of $\underset{\sim}{\tau}$ and $\underset{\sim}{\eta}$ as follows:

$$p(\underset{\sim}{\tau},\underset{\sim}{\eta}|\underset{\sim}{D},\underset{\sim}{U}) \approx p(\underset{\sim}{\tau},\underset{\sim}{\eta},\underset{\sim}{\xi} = \hat{\underset{\sim}{\xi}}_{\underset{\sim}{\tau},\underset{\sim}{\eta}}|\underset{\sim}{D},\underset{\sim}{U})|H|^{-1/2}$$

where

$$\underset{\sim}{H} = \left. \frac{\partial^2 \log p(\underset{\sim}{\xi},\underset{\sim}{\eta},\underset{\sim}{\tau})}{\partial \underset{\sim}{\xi}\ \partial \underset{\sim}{\xi}'} \right|_{\underset{\sim}{\xi} = \hat{\underset{\sim}{\xi}}_{\tau,\eta}} ,$$

with $\hat{\underset{\sim}{\xi}}_{\tau,\eta}$ denoting the modal value of $\underset{\sim}{\xi}$ from (2.2), evaluated at particular values of $\underset{\sim}{\tau}$ and $\underset{\sim}{\eta}$. In practice one would evaluate this expression at a grid of possible values of $\underset{\sim}{\tau}$ and $\underset{\sim}{\eta}$ in order to approximate their posterior marginal density, subsequently obtaining the mean and variance if desired. The approximation has the effect of replacing the integration in (2.3) with conditional maximizations, one for each point in the grid.

If item population parameters are not of interest, they can also be integrated out to yield

$$p(\underset{\sim}{\xi},\underset{\sim}{\tau}|\underset{\sim}{D},\underset{\sim}{U}) \propto \int_{\underset{\sim}{\eta}} \int_{\underset{\sim}{\theta}} p(\underset{\sim}{\theta},\underset{\sim}{\xi},\underset{\sim}{\tau},\underset{\sim}{\eta}|\underset{\sim}{D},\underset{\sim}{U})\ d\underset{\sim}{\theta}\ d\underset{\sim}{\eta} \quad . \tag{2.4}$$

The remaining item parameters and examinee population parameters are typically of primary interest in the educational setting, although for many examinees and all but very short tests, their marginal modes under (2.2) and (2.4) will differ little.

## An EM Algorithm for Parameter Estimation

This section provides a framework for parameter estimation in the general model outlined above, based on a variation of Dempster, Laird, and Rubin's EM algorithm introduced by Bock and Aitkin (1981) in the context of marginal maximum likelihood (MML) estimation of item parmeters. The posterior density function in our model, marginalized with respect to $\underset{\sim}{\theta}$, can be written as

$$p(\underset{\sim}{\xi},\underset{\sim}{\tau},\underset{\sim}{\eta}|\underset{\sim}{D},\underset{\sim}{U}) = \{\int_{\underset{\sim}{\theta}} L(\underset{\sim}{U}|\underset{\sim}{D},\underset{\sim}{\theta},\underset{\sim}{\xi})\ p(\underset{\sim}{\theta}|\underset{\sim}{\tau})\ d\underset{\sim}{\theta}\} \cdot \{p(\underset{\sim}{\tau})p(\underset{\sim}{\xi}|\underset{\sim}{\eta})p(\underset{\sim}{\eta})\} \quad . \tag{3.1}$$

The first bracketed expression on the right takes the form of the marginal likelihood of observed responses from a random sample of examinees from a population with density $p(\theta|\underset{\sim}{\tau})$, while the second can be thought of as the prior distribution for $\underset{\sim}{\xi}$ and $\underset{\sim}{\tau}$. We now focus our attention on the first term.

By maximizing the first term of (3.1) with respect to parameters of interest, Bock and Aitkin (1981) obtain MML estimates of $\underset{\sim}{\xi}$ given $p(\theta|\underset{\sim}{\tau})$ and Mislevy (1984) obtains MML estimates of $\underset{\sim}{\tau}$ given $L(\underset{\sim}{U}|\underset{\sim}{D},\underset{\sim}{\theta},\underset{\sim}{\xi})$. Both presentations employed the expedient of approximating integration over $\theta$ by summation over a finite grid of points $X_q$, $q = 1,\ldots,Q$, with associated weights $A(X_q|\underset{\sim}{\tau})$ as follows:

$$\log L(\underset{\sim}{U}|\underset{\sim}{D},\underset{\sim}{\xi},\underset{\sim}{\tau}) = \sum_i \log \sum_q L(\underset{\sim}{u}_i|\underset{\sim}{d}_i,X_q,\underset{\sim}{\xi})A(X_q) \quad . \tag{3.2}$$

Three methods were suggested for specifying points and weights. First, when $p(\theta|\underset{\sim}{\tau})$ takes the form of a normal density or a mixture of normal densities, optimal points and weights for a given Q may be found in Stroud and Sechrest (1966). Second, a Monte Carlo approach generates a random sample of equally-weighted points from $p(\theta|\underset{\sim}{\tau})$. Third, a grid of Q equally-spaced points can be specified a priori and assigned weights proportional to $p(X_q|\underset{\sim}{\tau})$.

Bock and Aitkin (1981) show that with the discrete approximation of the likelihood function, partial derivatives of the marginal likelihood, in which $\theta$'s are not observed but must be inferred from item responses, can be written in forms quite similar to their counterparts in a related "complete data" problem in which individual $\theta$'s are known. Under the assumption of iid $\theta$'s, we may write the partial derivative of the <u>complete data</u> log likelihood, namely

$$\log L(\underset{\sim}{U}|\underset{\sim}{D},\underset{\sim}{\theta},\underset{\sim}{\xi},\underset{\sim}{\tau}) = \log L(\underset{\sim}{U}|\underset{\sim}{D},\underset{\sim}{\theta},\underset{\sim}{\xi}) + \log p(\underset{\sim}{\theta}|\underset{\sim}{\tau}) \quad , \tag{3.3}$$

with respect to a typical parameter v from $\underset{\sim}{\xi}$ or $\underset{\sim}{\tau}$ in the form

$$\frac{\partial \log L(\underset{\sim}{U}|\underset{\sim}{D},\theta,\underset{\sim}{\xi},\underset{\sim}{\tau})}{\partial v} = \sum_i f_v(\underset{\sim}{r}_i,\underset{\sim}{N}_i,\theta_i,\underset{\sim}{\xi},\underset{\sim}{\tau}) \qquad (3.4)$$

for an appropriately defined gradient function $f_v$, where $N_{ij}$ is the number of attempts to item j by examinee i and $r_{ij}$ is the number of those that are correct. It can be shown (e.g., Mislevy, 1984) that the corresponding derivative of the <u>marginal</u> log likelihood (3.2) can then be approximated as

$$\frac{\partial \log L(\underset{\sim}{U}|\underset{\sim}{D},\underset{\sim}{\xi},\underset{\sim}{\tau})}{\partial v} = \sum_q f_v(\bar{\underset{\sim}{r}}_q,\bar{\underset{\sim}{N}}_q,X_q,\underset{\sim}{\xi},\underset{\sim}{\tau}) \quad , \qquad (3.5)$$

where

$$\bar{N}_{qj} = \sum_i d_{ij} P(X_q|\underset{\sim}{u}_i,\underset{\sim}{d}_i,\underset{\sim}{\xi},\underset{\sim}{\tau}) \qquad (3.6)$$

and

$$\bar{r}_{qj} = \sum_i d_{ij} u_{ij} P(X_q|\underset{\sim}{u}_i,\underset{\sim}{d}_i,\underset{\sim}{\xi},\underset{\sim}{\tau}) \qquad (3.7)$$

with

$$P(X_q|\underset{\sim}{u}_i,\underset{\sim}{d}_i,\underset{\sim}{\xi},\underset{\sim}{\tau}) = \frac{L(\underset{\sim}{u}_i|\underset{\sim}{d}_i,X_q,\underset{\sim}{\xi})A(X_q|\underset{\sim}{\tau})}{\sum_s L(\underset{\sim}{u}_i|\underset{\sim}{d}_i,X_q,\underset{\sim}{\xi})A(X_s|\underset{\sim}{\tau})} \quad . \qquad (3.8)$$

An application of Bayes theorem will be recognized in (3.8), yielding a value approximately proportional to the posterior density of $\theta$ given $\underset{\sim}{u}_i$, $\underset{\sim}{d}_i$, $\underset{\sim}{\xi}$, and $\underset{\sim}{\tau}$. The upshot is that the first derivatives (3.5) of the marginal likelihood are identical in form to the first derivatives (3.4) of the complete data likelihood, with expressions for subjects evaluated at $\theta_i$ with observed data $r_{ij}$ and $N_{ij}$ replaced by similar expressions evaluated at quadrature points $X_q$ with pseudo-data $\bar{r}_{qj}$ and $\bar{N}_{qj}$. Likelihood equations are obtained by setting the partial derivatives (3.5) to zero.

It will be noted that $\underset{\sim}{\bar{r}}_q$ and $\underset{\sim}{\bar{N}}_q$ depend on $\underset{\sim}{\xi}$ and $\underset{\sim}{\tau}$. Solution must proceed iteratively in EM cycles, which, with integration approximated by summation, take the form described by Dempster et al. (1977, Section 4.1.1) for missing values under multinomial sampling. In the E-step, (3.6) and (3.7) are evaluated with provisional estimates of $\underset{\sim}{\hat{\xi}}^t$ and $\underset{\sim}{\hat{\tau}}^t$. This gives the expectations of $\underset{\sim}{\bar{r}}_q$ and $\underset{\sim}{\bar{N}}_q$ conditional on the data and the provisional parameter estimates. In the M-step, $\underset{\sim}{\hat{\xi}}^{t+1}$ and $\underset{\sim}{\hat{\tau}}^{t+1}$ are obtained by solving (3.5) with with $\underset{\sim}{\bar{r}}_q$ and $\underset{\sim}{\bar{N}}_q$ treated as known. Cycles continue in this manner until changes become negligible. An indication of the precision

of estimation is given by the following approximation of the Fisher information matrix:

$$H = \sum_i \left( \frac{\partial \log L(\underset{\sim}{u}_i|\underset{\sim}{d}_i,\underset{\sim}{\xi},\underset{\sim}{\tau})}{\partial(\underset{\sim}{\xi},\underset{\sim}{\tau})} \right)\left( \frac{\partial \log L(\underset{\sim}{u}_i|\underset{\sim}{d}_i,\underset{\sim}{\xi},\underset{\sim}{\tau})}{\partial(\underset{\sim}{\xi},\underset{\sim}{\tau})'} \right) , \tag{3.9}$$

evaluated at $(\hat{\underset{\sim}{\xi}},\hat{\underset{\sim}{\tau}})$.

The EM algorithm is readily extended to Bayes modal estimation (Dempster et al., 1977, p. 6). All of the foregoing procedures are applied as before, except that the marginal likelihood equations (3.5) are replaced by so-called "Lindley equations"; for a typical element v of $\underset{\sim}{\xi}$ or $\underset{\sim}{\tau}$,

$$0 = \frac{\partial \log p(\underset{\sim}{\xi},\underset{\sim}{\tau}|\underset{\sim}{D},\underset{\sim}{U})}{\partial v} = \frac{\partial \log p(\underset{\sim}{U}|\underset{\sim}{D},\underset{\sim}{\xi},\underset{\sim}{\tau})}{\partial v} + \frac{\partial \log p(\underset{\sim}{\xi}|\underset{\sim}{\eta})}{\partial v}$$

$$+ \frac{\partial \log(\underset{\sim}{\tau})}{\partial v} . \tag{3.10}$$

The treatment of item population parameters $\underset{\sim}{\eta}$, which do not appear in (3.5), depends on whether they are to be integrated out or jointly estimated. Integrating them out modifies the form of the prior for $\underset{\sim}{\xi}$ from $p(\underset{\sim}{\xi}|\underset{\sim}{\eta})$ to $\int p(\underset{\sim}{\xi}|\underset{\sim}{\eta})p(\underset{\sim}{\eta})\, d\underset{\sim}{\eta}$. Estimating them requires the solution of additional equations

$$0 = \frac{\partial \log p(\underset{\sim}{\xi}|\underset{\sim}{\eta})p(\underset{\sim}{\eta})}{\partial \underset{\sim}{\eta}} \quad . \tag{3.11}$$

Under regularity conditions, posterior densities in Bayes estimation tend to multivariate normality as sample size increases. Asymptotically, the mean is equal to the mode, which is equal to the maximum likelihood estimate. The precision matrix, or the inverse of the covariance matrix, is given by the negative matrix of second derivatives of the log posterior, evaluated at that point.

When $\underset{\sim}{\eta}$ has been integrated out in the problem at hand, this matrix takes the form

$$\underset{\sim}{A} = -\frac{\partial^2 \log L(\underset{\sim}{U}|\underset{\sim}{D},\underset{\sim}{\xi},\underset{\sim}{\tau})}{\partial(\underset{\sim}{\xi},\underset{\sim}{\tau})\,\partial(\underset{\sim}{\xi},\underset{\sim}{\tau})'} - \frac{\partial^2 \log p(\underset{\sim}{\xi})p(\underset{\sim}{\tau})}{\partial(\underset{\sim}{\xi},\underset{\sim}{\tau})\,\partial(\underset{\sim}{\xi},\underset{\sim}{\tau})'} \tag{3.12}$$

where

$$p(\underset{\sim}{\xi}) = \int p(\underset{\sim}{\xi}|\underset{\sim}{\eta})\, p(\underset{\sim}{\eta})\, d\underset{\sim}{\eta} \quad .$$

Employing the well-known result on Fisher's information matrix

$$E\left(-\frac{\partial^2 \log(\text{data}|\underset{\sim}{x})}{\partial \underset{\sim}{x}\,\partial \underset{\sim}{x}'}\right) = E\left[\left(\frac{\partial \log(\text{data}|\underset{\sim}{x})}{\partial \underset{\sim}{x}}\right)\left(\frac{\partial \log(\text{data}|\underset{\sim}{x})}{\partial \underset{\sim}{x}'}\right)\right]$$

(Kendall & Stuart, 1973, pp. 8-10) and substituting observed values for expectations, we avoid calculating second derivatives of the log likelihood via the approximation

$$\underset{\sim}{A} \approx -H - \frac{\partial \log p(\underset{\sim}{\xi})p(\underset{\sim}{\tau})}{\partial(\underset{\sim}{\xi},\underset{\sim}{\tau})\ \partial(\underset{\sim}{\xi},\underset{\sim}{\tau})'} \quad . \tag{3.13}$$

where H is given in (3.9). When $\underset{\sim}{\eta}$ is estimated jointly with $\underset{\sim}{\xi}$ and $\underset{\sim}{\tau}$, the precision matrix is similarly approximated as

$$\underset{\sim}{B} = \begin{vmatrix} -H - \dfrac{\partial^2 \log p(\underset{\sim}{\tau})p(\underset{\sim}{\xi}|\underset{\sim}{\eta})}{\partial(\underset{\sim}{\xi},\underset{\sim}{\tau})\ \partial(\underset{\sim}{\xi},\underset{\sim}{\tau})'} & \text{(symmetric)} \\ -\dfrac{\partial^2 \log p(\underset{\sim}{\tau})p(\underset{\sim}{\xi}|\underset{\sim}{\eta})p(\underset{\sim}{\eta})}{\partial(\underset{\sim}{\xi},\underset{\sim}{\tau})\ \partial\underset{\sim}{\eta}'} & -\dfrac{\partial^2 \log \underset{\sim}{p}(\underset{\sim}{\tau})\underset{\sim}{p}(\underset{\sim}{\xi}|\underset{\sim}{\eta})\underset{\sim}{p}(\underset{\sim}{\eta})}{\partial\underset{\sim}{\eta}\ \partial\underset{\sim}{\eta}'} \end{vmatrix} . \tag{3.14}$$

It should be pointed out that solutions of the Lindley equations are local extrema or saddle points of the posterior. Whether they are local maxima can be determined by examining the shape of the posterior in the neighborhoods of solutions, either empirically or though the matrix of second derivatives, which will be negative definite at local maxima. Whether a local maximum is a global maximum follows in certain cases from the form of the posterior

(e.g., a member of the exponential family), but must be determined empirically in most cases by starting the iterative solution from a number of different initial values.

## Procedures for Some Logistic Models

The balance of the article implements the procedures in the context of logistic item response models. The following sections provide details on functional forms for the likelihood and prior distributions, and on the corresponding forms of the fitting equations. For the first stage of priors, a multivariate normal density will be posited for item thresholds, log slopes, and logit asymptotes; both a mixture of normal components and a nonparametric approximation in the form of a histogram will be provided for examinee abilities. For the second stage, both diffuse and natural conjugate priors will be provided in all cases.

## The Likelihood Function

The three-parameter logistic model for dichotomous items (Birnbaum, 1968) gives the probability of a correct response to item j from examinee i as

$$P_j(\theta_i) = P(u_{ij} = 1|\theta_i, a_j, b_j, c_j)$$

$$= c_j + (1 - c_j)\Psi[Da_j(\theta_i - b_j)] \quad , \qquad (4.1)$$

where $\Psi(x)$ is the logistic function $1/(1 + \exp(-x))$. D is a scaling constant, taken as 1 by some writers for convenience and as 1.7 by others (e.g., Birnbaum, 1968) so that the units of the model will approximate those of normal ogive IRT models (Lord, 1952). One may obtain the two-parameter logistic model from (4.1) by fixing $c_j = 0$, and the one-parameter model (Rasch, 1960) by additionally fixing $a_j = 1$.

Indeterminacies of scale and origin are apparent in (4.1). If for any scalars m and x we define $\theta^* = m\theta + x$, $b^* = mb + x$, and $a^* = a/m$, then $P(u = 1|\theta^*,a^*,b^*,c) = P(u = 1|\theta,a,b,c)$. In this article we will specify higher-level prior distributions that resolve these indeterminacies.

Rather than obtaining a posterior for $\underset{\sim}{a}$, $\underset{\sim}{b}$, and $\underset{\sim}{c}$ directly, we work with the transformed item parameters

$$\alpha_j = \log a_j$$

$$\beta_j = b_j$$

and

$$\gamma_j = \log(c_j/(1 - c_j)) \quad .$$

It is readily inferred that $a_j = \exp \alpha_j$ and $c_j = \Psi(\gamma_j)$. While this formulation does not permit the boundary values of 0 and 1 for $c_j$, it serves our purposes adequately by allowing c's arbitrarily close to these values. Non-positive a's are also disallowed; careful examination of fitted and empirical response curves will obviously be required in applications where faulty items and incorrect keys can occur.

Reparameterization achieves two ends. The first is a more rapid attainment of large-sample results. The impediment against normality represented by the finite range of c, for example, is removed by re-expression in terms of $\gamma$. The second is convenience in specifying higher level prior densitites. With unrestricted ranges for all parameters, the imposition of multivariate normal priors on parameters within items but independent across items is not unreasonable. This may be the simplest way to allow for the possibility of dependence among parameters a, b, and c in a population of items.

Letting $\underset{\sim}{\xi}$ represent $(\alpha_1, \beta_1, \gamma_1, \ldots, \alpha_n, \beta_n, \gamma_n)$, the Lindley equations for item parameters take the form

$$0 = \frac{\partial \log L(\underset{\sim}{\xi}, \underset{\sim}{\tau})}{\partial \underset{\sim}{\xi}} + \frac{\partial \log p(\underset{\sim}{\xi} | \underset{\sim}{\eta})}{\partial \underset{\sim}{\xi}} \quad . \tag{4.2}$$

Formulas for the second term appear in the following section. Those for the first term are approximated as

$$\frac{\partial \log L}{\partial \alpha_j} = D\,(\exp \alpha_j)(1 - c_j) \sum_q (\bar{r}_{qj} - \bar{N}_{qj} P_{qj}) W_{qj} (X_k - b_j) \quad , \tag{4.3}$$

$$\frac{\partial \log L}{\partial \beta_j} = -D(1 - c_j) \sum_q (\bar{r}_{qj} - \bar{N}_{qj} P_{qj}) W_{qj} a_j \quad , \tag{4.4}$$

$$\frac{\partial \log L}{\partial \gamma_j} = c_j \sum_q (\bar{r}_{qj} - \bar{N}_{qj} P_{qj}) / P_{qj} \quad , \tag{4.5}$$

where $\bar{N}_{qj}$ and $\bar{r}_{qj}$ are given in (3.6) and (3.7) and

$$W_{qj} = [P^*_{qj}(1 - P^*_{qj})]/[P_{qj}(1 - P_{qj})]$$

with

$$P_{qj} = c_j + (1 - c_j)\Psi[Da_j(X_k - b_j)]$$

and

$$P^*_{qj} = \Psi[Da_j(X_k - b_j)] \quad .$$

Given $\bar{N}_{qj}$ and $\bar{r}_{qj}$, the equations (4.2) corresponding to parameters of different items are independent. This means that the M-step task of finding zeros of (4.2), along with additional

Lindley equations for examinee- and possibly item-population parameters, need not address all 3n equations for item parameters simultaneously. Zeros for the parameters of a given item may be obtained rapidly by methods such as Newton-Raphson iterations, which require second derivatives of the log posterior, or Davidon-Fletcher-Powell iterations, which do not.

## Structures on Item Parameters

Let the prior distribution on the parameters for item j be given by $\underset{\sim}{\xi}_j = (\alpha_j, \beta_j, \gamma_j) \sim \text{MVN}(\underset{\sim}{\mu}_\xi, \underset{\sim}{\Sigma}_\xi)$, where $\underset{\sim}{\mu}_\xi = (\mu_\alpha, \mu_\beta, \mu_\gamma)$. Hence $(\underset{\sim}{\mu}_\xi, \underset{\sim}{\Sigma}_\xi)$ plays the role of the item population parameter $\underset{\sim}{\eta}$ in the more general notation of the preceding section. Assuming independence over items, the joint prior density of item parameters is then given by

$$p(\underset{\sim}{\xi} \mid \underset{\sim}{\mu}_\xi, \underset{\sim}{\Sigma}_\xi) \propto |\underset{\sim}{\Sigma}_\xi|^{-n/2} \prod_j \exp\{-\tfrac{1}{2} (\underset{\sim}{\xi}_j - \underset{\sim}{\mu}_\xi)' \underset{\sim}{\Sigma}_\xi^{-1} (\underset{\sim}{\xi}_j - \underset{\sim}{\mu}_\xi)\} \qquad (5.1)$$

and the log prior density by

$$\log p(\underset{\sim}{\xi} \mid \underset{\sim}{\mu}_\xi, \underset{\sim}{\Sigma}_\xi) = -\frac{n}{2} \log |\underset{\sim}{\Sigma}_\xi| - \frac{1}{2} \sum_j (\underset{\sim}{\xi}_j - \underset{\sim}{\mu}_\xi)' \underset{\sim}{\Sigma}_\xi^{-1} (\underset{\sim}{\xi}_j - \underset{\sim}{\mu}_\xi) \quad . \qquad (5.2)$$

The partial derivatives of (5.2) with respect to the parameters for item j are obtained as

$$\frac{\partial \log p(\underset{\sim}{\xi} | \underset{\sim}{\mu}_{\xi}, \underset{\sim}{\Sigma}_{\xi})}{\partial \underset{\sim}{\xi}_j} = -\underset{\sim}{\Sigma}_{\xi}^{-1}(\underset{\sim}{\xi}_j - \underset{\sim}{\mu}_{\xi})$$

$$= \begin{vmatrix} -\sigma_{\xi}^{11}(\alpha_j - \mu_{\alpha}) - \sigma_{\xi}^{12}(\beta_j - \mu_{\beta}) - \sigma_{\xi}^{13}(\gamma_j - \mu_{\gamma}) \\ -\sigma_{\xi}^{21}(\alpha_j - \mu_{\alpha}) - \sigma_{\xi}^{22}(\beta_j - \mu_{\beta}) - \sigma_{\xi}^{23}(\gamma_j - \mu_{\gamma}) \\ -\sigma_{\xi}^{31}(\alpha_j - \mu_{\alpha}) - \sigma_{\xi}^{32}(\beta_j - \mu_{\beta}) - \sigma_{\xi}^{33}(\gamma_j - \mu_{\gamma}) \end{vmatrix} . \quad (5.3)$$

These terms are added to the partial derivatives of the log likelihood (4.3)-(4.5) and the results set to zero to give the Lindley equations for the parameters of item j.

In IRT models with independent unimodal prior distributions on item parameters, the contribution of prior information in the Lindley equation for a given parameter depends upon its distance from the center of the distribution of parameters of its same type. That is, parameters of a given type "shrink" toward a single point, namely the mean of parameters of that type, by amounts inversely proportional to the information available each individually.

It will be seen in (5.3) that under the structure proposed here, the contribution of the prior also depends on the distance of the item's parameters of other types from the centers of their respective populations. Parameters of a given type now shrink toward a plane, namely their conditional expectations given the values of the items' parameters of other types.

Let us suppose further that $(\underset{\sim}{\mu}_\xi, \underset{\sim}{\Sigma}_\xi)$ follows the natural conjugate prior distribution for the mutlivariate normal, namely multivariate normal for $\underset{\sim}{\mu}_\xi$ given $\underset{\sim}{\Sigma}_\xi$ and inverted Wishart for $\underset{\sim}{\Sigma}_\xi$ (Ando and Kaufman, 1965):

$$p(\underset{\sim}{\mu}_\xi, \underset{\sim}{\Sigma}_\xi) \propto |\underset{\sim}{\Sigma}_\xi^{-1}|^{(m+1)/2} \exp\{-\tfrac{1}{2}[(\underset{\sim}{\mu}_\xi - \underset{\sim}{y})'\underset{\sim}{\Sigma}_\xi^{-1}(\underset{\sim}{\mu}_\xi - \underset{\sim}{y})b + \operatorname{tr} \underset{\sim}{\Sigma}_\xi^{-1}\underset{\sim}{H}]\} \tag{5.4}$$

whence

$$\log p(\underset{\sim}{\mu}_\xi, \underset{\sim}{\Sigma}_\xi) \propto -(m+1)/2 \log|\underset{\sim}{\Sigma}_\xi| - \tfrac{1}{2}(\underset{\sim}{\mu}_\xi - \underset{\sim}{y}_\xi)\underset{\sim}{\Sigma}_\xi^{-1}(\underset{\sim}{\mu}_\xi - \underset{\sim}{y}_\xi)b - \tfrac{1}{2}\operatorname{tr} \underset{\sim}{\Sigma}_\xi^{-1}\underset{\sim}{H} \quad . \tag{5.5}$$

Here b and m is a scalars ($m > 2p$ for a proper distribution under the p-parameter IRT model), $\underset{\sim}{y}_\xi$ is a vector, and $\underset{\sim}{H}$ is a 3-by-3 positive symmetric matrix—all to be specified in such a way as to $\underset{\sim}{H}$ corresponds to the covariance of m - p values of $\underset{\sim}{\xi}$ and $\underset{\sim}{y}_\xi$ corresponds to the average of the b values of $\underset{\sim}{\xi}$.

The indeterminacies of scale and origin in the two- and three-parameter models can be conveniently resolved at this point by specifying that $p(\underset{\sim}{\mu}_\xi, \underset{\sim}{\Sigma}_\xi)$ is null everywhere except where $\mu_\alpha = 0$ and $\mu_\beta = 0$. Only the latter constraint enters into the one-parameter model.

If $\underset{\sim}{\mu}_\xi$ and $\underset{\sim}{\Sigma}_\xi$ are to be estimated jointly with $\underset{\sim}{\xi}$ and $\underset{\sim}{\tau}$, partial derivatives must first be obtained for all terms in the log posterior in which they appear, namely $\log p(\underset{\sim}{\xi} | \underset{\sim}{\mu}_\xi, \underset{\sim}{\Sigma}_\xi)$ (5.2) and $\log p(\underset{\sim}{\mu}_\xi, \underset{\sim}{\Sigma}_\xi)$ (5.5):

$$\frac{\partial \log p(\underset{\sim}{\xi} | \underset{\sim}{\mu}_\xi, \underset{\sim}{\Sigma}_\xi)}{\partial \underset{\sim}{\mu}_\xi} + \frac{\partial \log p(\underset{\sim}{\mu}_\xi, \underset{\sim}{\Sigma}_\xi)}{\partial \underset{\sim}{\mu}_\xi}$$

$$= \frac{1}{2} \underset{\sim}{\Sigma}_\xi^{-1} \{\Sigma_j (\underset{\sim}{\xi}_j - \underset{\sim}{\mu}_\xi) - b(\underset{\sim}{\mu}_\xi - \underset{\sim}{y}_\xi)\} \qquad (5.6)$$

and

$$\frac{\partial \log p(\underset{\sim}{\xi} | \underset{\sim}{\mu}_\xi, \underset{\sim}{\Sigma}_\xi)}{\partial \underset{\sim}{\Sigma}_\xi} + \frac{\partial \log p(\underset{\sim}{\mu}_\xi, \underset{\sim}{\Sigma}_\xi)}{\partial \underset{\sim}{\Sigma}_\xi}$$

$$= \frac{1}{2} \underset{\sim}{\Sigma}_\xi^{-1} - (n + m + 1) \underset{\sim}{\Sigma}_\xi + \underset{\sim}{S} + n(\bar{\underset{\sim}{\xi}} - \underset{\sim}{\mu}_\xi)(\bar{\underset{\sim}{\xi}} - \underset{\sim}{\mu}_\xi)'$$

$$+ b(\underset{\sim}{\mu}_\xi - \underset{\sim}{y}_\xi)(\underset{\sim}{\mu}_\xi - y_\xi)' + \underset{\sim}{H}\}\underset{\sim}{\Sigma}_\xi^{-1} \qquad (5.7)$$

where

$$\bar{\underset{\sim}{\xi}} = n^{-1} \sum_j \underset{\sim}{\xi}_j \qquad (5.8)$$

and

$$\underset{\sim}{S} = \sum_j (\underset{\sim}{\xi}_j - \bar{\underset{\sim}{\xi}})(\underset{\sim}{\xi}_j - \bar{\underset{\sim}{\xi}})' \quad . \qquad (5.9)$$

Equating to zero and simplifying yields the Lindley equations

$$\underset{\sim}{\mu}_{\xi} = \frac{n\bar{\underset{\sim}{\xi}} + b\underset{\sim}{y}_{\xi}}{n + b} \tag{5.10}$$

and

$$\underset{\sim}{\Sigma}_{\xi} = (n + m + 1)^{-1}\{\underset{\sim}{S} + n(\bar{\underset{\sim}{\xi}} - \underset{\sim}{\mu}_{\xi})(\bar{\underset{\sim}{\xi}} - \underset{\sim}{\mu}_{\xi})'$$

$$+ b(\underset{\sim}{\mu}_{\xi} - \underset{\sim}{y}_{\xi})(\underset{\sim}{\mu}_{\xi} - \underset{\sim}{y}_{\xi})' + \underset{\sim}{H}\} \quad . \tag{5.11}$$

A familiar theme in Bayesian estimation appears in (5.10), where a mean is estimated as a weighted average of a sample mean and a prior mean. It should be pointed out that $\bar{\underset{\sim}{\xi}}$ in (5.8) will generally not be equal to the simple mean of the item parameter estimates that would have been obtained under joint maximum likelihood (JML) estimation. This is because the item parameters $\xi_j$ are being estimated at the same time, and each is shrinking back from its JML value in <u>inverse</u> proportion to the amount of information about it; items therefore contribute toward the estimation of the item population mean in <u>direct</u> proportion to the amount of information about them.

To emulate maximum likelihood estimation of $\underset{\sim}{\mu}_\xi$ and $\underset{\sim}{\Sigma}_\xi$, again jointly with $\underset{\sim}{\xi}$ and $\underset{\sim}{\tau}$, one may specify that $\underset{\sim}{H} = 0$ and $m = 2$, and omit the quadratic term involving $\underset{\sim}{y}_\xi$ in and after (5.4). This gives an improper diffuse prior, justifyable along the lines of invariance with respect to reparameterization (Jeffreys, 1961). The partial derivatives and Lindley equations simplify in obvious ways.

If modal values of $\underset{\sim}{\xi}$ and $\underset{\sim}{\tau}$ marginalized with respect to $\underset{\sim}{\mu}_\xi$ and and $\underset{\sim}{\Sigma}_\xi$ are desired, these latter parameters may be integrated out and then Lindley equations for item parameters modified in the following manner. Focusing on the relevant terms of the posterior, we can write

$$p(\underset{\sim}{\xi} | \underset{\sim}{\mu}_\xi, \underset{\sim}{\Sigma}_\xi) \cdot p(\underset{\sim}{\mu}_\xi, \underset{\sim}{\Sigma}_\xi)$$

$$\propto |\underset{\sim}{\Sigma}_\xi|^{-(n+m+1)/2} \exp\{ -\frac{1}{2} \operatorname{tr} \underset{\sim}{\Sigma}_\xi^{-1} [\underset{\sim}{S} + \underset{\sim}{H} + n(\bar{\underset{\sim}{\xi}} - \underset{\sim}{\mu}_\xi)(\bar{\underset{\sim}{\xi}} - \underset{\sim}{\mu}_\xi)'$$

$$+ b(\underset{\sim}{\mu}_\xi - \underset{\sim}{y}_\xi)(\underset{\sim}{\mu}_\xi - \underset{\sim}{y}_\xi)']\} \quad . \tag{5.12}$$

Integration over $\underset{\sim}{\Sigma}_\xi$ yields a multivariate-t distribution for $\underset{\sim}{\mu}_\xi$ (Ando and Kaufman, 1965):

$$p(\underset{\sim}{\xi}|\underset{\sim}{\mu}_{\xi})p(\underset{\sim}{\mu}_{\xi}) \propto [1 + (\underset{\sim}{\mu}_{\xi} - \underset{\sim}{v})'C^{-1}(\underset{\sim}{\mu}_{\xi} - \underset{\sim}{v})]^{-(n+m-p)/2}$$

where

$$\underset{\sim}{v} = \frac{b\underset{\sim}{y}_{\xi} + n\bar{\underset{\sim}{\xi}}}{b + n}$$

and

$$\underset{\sim}{C} = \frac{\underset{\sim}{S} + \underset{\sim}{H}}{n + l} + \frac{nb}{(n + b)^2} (\underset{\sim}{y}_{\xi} - \bar{\underset{\sim}{\xi}})(\underset{\sim}{y}_{\xi} - \bar{\underset{\sim}{\xi}})' \cdot .$$

By using the constant of integration for the multivariate-t, we obtain for the marginal distribution of $\underset{\sim}{\xi}$ the following quantity:

$$p(\underset{\sim}{\xi}) \propto |\underset{\sim}{C}|^{1/2} .$$

The terms to be added to the partial derivative of the log marginal likelihood to obtain a Lindley equation for $\underset{\sim}{\xi}_j$, now marginalized with respect to $\underset{\sim}{\mu}_{\xi}$ and $\underset{\sim}{\Sigma}_{\xi}$, become

$$\frac{\partial \log p(\underset{\sim}{\xi})}{\partial \underset{\sim}{\xi}_j} = - \frac{\underset{\sim}{C}^{-1}}{n+1} \left[ \xi_j - \bar{\underset{\sim}{\xi}} + \frac{\bar{\underset{\sim}{\xi}} - \underset{\sim}{y}_{\xi}}{n+1} \right] .$$

This result is similar in form to (5.3), the contribution when $\underset{\sim}{\mu}_{\xi}$ and $\underset{\sim}{\Sigma}_{\xi}$ are estimated jointly with $\underset{\sim}{\xi}$.

## Structures on Examinee Parameters

This section presents details for two types of prior distributions on examinee parameters $\underset{\sim}{\tau}$, namely a nonparameteric prior in the form of a histogram and a mixture of homoscedastic normal distributions in unknown proportions. The latter choice includes the familiar standard normal prior as a special case.

Recalling the form of the posterior distribution for $\underset{\sim}{\xi}$, $\underset{\sim}{\eta}$, and $\underset{\sim}{\tau}$, or

$$p(\underset{\sim}{\xi},\underset{\sim}{\tau},\underset{\sim}{\eta}|\underset{\sim}{D},\underset{\sim}{U}) = \{ \int_{\underset{\sim}{\theta}} L(\underset{\sim}{U}|\underset{\sim}{D},\underset{\sim}{\theta},\underset{\sim}{\xi}) p(\underset{\sim}{\theta}|\underset{\sim}{\tau}) \, d\underset{\sim}{\theta} \} \cdot \{p(\underset{\sim}{\tau})\} \cdot \{p(\underset{\sim}{\xi}|\underset{\sim}{\eta}) p(\underset{\sim}{\eta})\} , \tag{6.1}$$

we note that (1) contributions to the Lindley equations for $\underset{\sim}{\tau}$ come from the marginal likelihood and from its prior and (2) these contributions are the same regardless of whether $\underset{\sim}{\eta}$ is being

estimated jointly or integrated out. Both partial derivatives and Lindley equations for $\underset{\sim}{\tau}$ are presented here, the former because they are needed to approximate the information matrix and the latter because the partial derivatives often simplify after being equated to zero. Detailed calculations of the contributions from the marginal likelihood are omitted, as they may be found in Mislevy (1984).

A nonparametric solution: If $p(\underset{\sim}{\theta}|\underset{\sim}{\tau})$ is a smooth continuous density, it may be approximated by a discrete distribution over a finite number of points $X_q$, $q = 1,\ldots,Q$. Letting $p_q$ denote the density at point $X_q$, we approximate the log marginal likelihood as

$$\log L(\underset{\sim}{U}|\underset{\sim}{D},\underset{\sim}{\xi},\underset{\sim}{\tau}) = \sum_{i=1}^{N} \log h(\underset{\sim}{u}_i) \qquad (6.2)$$

where

$$h(\underset{\sim}{u}_i) = \sum_{q=1}^{Q} L(\underset{\sim}{u}_i|\underset{\sim}{d}_i, X_q, \underset{\sim}{\xi}) p_q \quad .$$

The continuous density $p(\underset{\sim}{\theta}|\underset{\sim}{\tau})$ is thus replaced by a multinomial distribution with parameters $p_1,\ldots,p_{Q-1}$, with

$$p_Q = 1 - \sum_{q=1}^{Q-1} p_q \quad .$$

It can be shown that the partial derivative of (6.2) with respect to $p_q$ is

$$\frac{\partial \log L}{\partial p_q} = \sum_i h^{-1}(\underset{\sim}{u}_i)[L(\underset{\sim}{u}_i|\underset{\sim}{d}_i, X_q, \underset{\sim}{\xi}) - L(\underset{\sim}{u}_i|\underset{\sim}{d}_i, X_Q, \underset{\sim}{\xi})] \quad .$$

The natural conjugate prior for the multinomial is the Dirchlet distribution, which takes the following form:

$$p(p_1,\ldots,p_{Q-1}|M_1,\ldots,M_Q) \propto \prod_k p_k^{M_k - 1} \quad ,$$

which implies that

$$\frac{\partial \log p(\underset{\sim}{p}|\underset{\sim}{M})}{\partial p_q} = \frac{M_q - 1}{p_q} - \frac{M_Q - 1}{p_Q} \quad , \qquad q = 1,\ldots Q-1 \quad .$$

Prior belief about $p_1,\ldots,p_Q$ are thus expressed as values of the proportions $(M_1 - 1)/M^+,\ldots,(M_Q - 1)/M^+$, where $M^+ = \Sigma M_q - 1$.

The forms given above provide first derivatives that lead to a positive definite matrix of second derivatives, and are thereby useful in estimating parameters by Newton or quasi-Newton algorithms and in computing posterior variances. Simpler and more intuitively appealing Lindley equations result, however, for all Q p's with their sum restricted to unity:

$$p_q = \frac{\bar{N}_q + (M_q - 1)}{N + M^+} \quad , \qquad q = 1,\ldots,Q \quad , \tag{6.4}$$

where

$$\bar{N}_q = \sum_i p(X_q | \underset{\sim}{u}_i, \underset{\sim}{d}_i, \underset{\sim}{\xi}, \underset{\sim}{p})$$

$$= \frac{\sum_i L(\underset{\sim}{u}_i | \underset{\sim}{d}_i, X_q, \underset{\sim}{\xi}, \underset{\sim}{p}) p_q}{\sum_r \sum_i L(\underset{\sim}{u}_i | \underset{\sim}{d}_i, X_r, \underset{\sim}{\xi}, \underset{\sim}{p}) p_r} \quad .$$

The posterior density at point $X_q$, therefore, is a weighted average of its prior density and the expectation of its density conditional on the data and the densities themselves.

To obtain maximum likelihood estimates, we may take a uniform diffuse prior with $M_q \equiv 1$. An alternative diffuse prior with

$M_q \equiv -1$ may be preferred, however, on the grounds of robustness with respect to the choice of quadrature points (Novick & Jackson, 1974, p. 347 ff.).

It is possible to resolve the indeterminacies of the IRT model at the point, by specifying that the distribution $p(\underset{\sim}{p}|\underset{\sim}{M})$ can take nonzero values only when the following equality constraints are satisfied:

$$\sum_q X_q p_q = 0$$

and

$$\sum_q X_q^2 p_q = 1 \quad .$$

Values of $\underset{\sim}{M}$ specified in an informative prior should satisfy these constraints as well.

A mixture of normal components. Suppose that the distribution is a mixture of K normal components, with means $\underset{\sim}{\mu} = (\mu_1,\ldots,\mu_K)$ and common variance $\sigma^2$. Let $\underset{\sim}{p} = (p_1,\ldots,p_K)$ be the unknown proportions of the mixture. Define the marginal probability of response pattern $\underset{\sim}{u}$ given $\underset{\sim}{\xi}$ and $\underset{\sim}{\tau} = (\underset{\sim}{\mu},\underset{\sim}{p},\sigma^2)$ as

$$h(\underset{\sim}{u}) = \sum_k p_k \int_\theta L(\underset{\sim}{u}|\underset{\sim}{d},\theta,\underset{\sim}{\xi},\underset{\sim}{\tau}) f_k(\theta)\, d\theta \quad ,$$

where

$$f_k(\theta) = \frac{1}{\sqrt{2\pi}\,\sigma} \exp\left[ \frac{-(\theta - \mu_k)^2}{2\sigma^2} \right] \quad .$$

The log marginal likelihood for N examinees is then written as

$$\log L(\underset{\sim}{U}|\underset{\sim}{D},\underset{\sim}{\xi},\underset{\sim}{\tau}) = \sum_i \log h(\underset{\sim}{u}_i) \quad . \tag{6.5}$$

Approximating integration by summation over a fixed grid of equally-spaced quadrature points $X_1,\ldots,X_Q$,

$$\log L(\underset{\sim}{U}|\underset{\sim}{D},\underset{\sim}{\xi},\underset{\sim}{\tau}) = \sum_i \log \sum_k p_k \sum_q L(\underset{\sim}{u}_i|X_q) f_k(X_q) \quad ,$$

where

$$L(\underset{\sim}{u}_i|X_q) = L(\underset{\sim}{u}_i|\underset{\sim}{d}_i,X_q,\underset{\sim}{\xi},\underset{\sim}{\tau}) \quad .$$

Taking $p_1,\ldots,p_{Q-1}$ as the parameters specifying proportions, partial derivatives of (6.5) are then obtained as

$$\frac{\partial \log L}{\partial p_k} = p_k^{-1} \sum_q \bar{N}_{kq} - p_K^{-1} \sum_q \bar{N}_{Kq} \quad , \quad k = 1,\ldots,Q-1 \quad , \tag{6.6}$$

$$\frac{\partial \log L}{\partial \mu_k} = \sigma^{-2} \sum_q \bar{N}_{kq}(X_q - \mu_k) \quad , \tag{6.7}$$

and

$$\frac{\partial \log L}{\partial \sigma^2} = -\frac{N}{2\sigma^2} + \frac{1}{2\sigma^4} \sum_k \sum_q \bar{N}_{kq}(X_q - \mu_k)^2 \quad , \tag{6.8}$$

where

$$\bar{N}_{kq} = \sum_i h^{-1}(\underset{\sim}{u}_i) p_k L(\underset{\sim}{u}_i | X_q) f_k(X_q) \quad .$$

A natural conjugate prior for $\underset{\sim}{\tau}$ is Dirichlet-normal-inverse gamma:

$$p(\underset{\sim}{p},\underset{\sim}{\mu},\sigma^2) \propto \{ \prod_k p_k^{M_k - 1} \} \{ \prod_k \exp [ \frac{-(\mu_k - y_k)^2}{2\sigma^2} ] \}$$

$$\{\sigma^{-(v/2+1)} \exp ( \frac{-s}{2\sigma^2} \} \quad , \qquad (6.9)$$

whence

$$\log p(\underset{\sim}{p},\underset{\sim}{\mu},\sigma^2) = \sum_k (M_k - 1) \log p_k + \sum_k \frac{-(\mu_k - y_k)^2}{2\sigma^2}$$

$$- (v/2 + 1) \log \sigma - (s/2\sigma^2) \quad . \qquad (6.10)$$

Here $\underset{\sim}{M}$, $\underset{\sim}{y}$, v, and s are the parameters of the prior distribution, to be supplied by the user. $\underset{\sim}{M}$ can be thought of as the number of examinees in each of the components from a sample of size $M^+ = \sum M_k - 1$; $\underset{\sim}{y}$ can be thought of as anticipated locations for the means of the components. v and s are the parameters of the inverted gamma distribution, possibly more easily specified after

one has in mind a mean and variance of such a distribution that incorporates prior belief about $\sigma^2$:

$$v = \frac{2\ \text{mean}^2}{\text{variance}} + 4$$

and

$$s = \frac{\text{mean} \cdot \text{variance}}{2(\text{mean} + \text{variance}^2)} \quad .$$

The indeterminacies of the IRT model can also be resolved at this point, by specifying that the total mean and within-component variance take specified values, say 0 and 1. That is, $p(\underset{\sim}{\tau})$ is zero except where

$$\sum_k p_k \mu_k = 0$$

and

$$\sigma^2 = 1 \quad .$$

When K = 1, a standard normal density is effectively specified for $\theta$ by this procedure.

Lindley equations are now obtained as the sums of partial derivatives of the log marginal likelihood (6.5) and the log prior (6.10). Again writing equations in terms of K p's constrained to a sum of one, we obtain

$$p_k = \frac{\sum_q \bar{N}_{kq} + (M_k - 1)}{N + M^+} \quad , \qquad k = 1,,\ldots,Q \quad , \tag{6.11}$$

$$\mu_k = \frac{\sum_q \bar{N}_{kq} X_q + y_k}{\sum_q \bar{N}_{kq} + 1} \tag{6.12}$$

and

$$\sigma^2 = \frac{\sum_k \sum_q \bar{N}_{kq} (X_q - \mu_k)^2 + \sum_k (\mu_k - y_k)^2 + s}{N + (v/2 + 1)} \quad . \tag{6.13}$$

A diffuse prior may be obtained from (6.9) by omitting the term involving $\underset{\sim}{\mu}$ and setting $M_k \equiv 1$, s = 0, and v = 0. Partial derivatives and Lindley equations simplify in obvious ways.

## A Numerical Example

Satisfactory procedures for item parameter estimation have been available for some time for both large and small samples under the one-parameter logistic (Rasch) IRT model and for large samples of both persons and items under the three-parameter logistic (Birnbaum) IRT models. The same cannot be said about small samples under the three-parameter model, and it is to this problem we apply the procedures of the proceeding sections.

A perusal of the recent literature on Bayesian item parameter estimation suggests that such efforts were motivated not so much by the pursuit of minimum mean squared error or by a conviction that all unknowns should be expressed in probabilistic terms, but rather by a more practical desire to obtain "reasonable" item parameter estimates—in particular, finite ones.

The essential difficulty with parameter estimation under the three-parameter model is that the parameters of a given item are often poorly determined by the data at hand; apparently discrepant triples (a, b, c) can trace similar response curves in the region of the ability scale where the sample of examinees is to be found. Such poor resolution is manifest as a likelihood surface nearly flat along one or more dimensions, yielding unstable maximum likelihood estimates (MLE's). A trivially higher likelihood may be produced, for example, by taking a particular item's values of a and c to be 200 and .6 rather than the more reasonable values of 2 and .25.

Extreme and infinite parameter estimates can be avoided by using a single-stage Bayesian prior, but not without introducing an additional hazard. A fully-specified prior will indeed have the desired effect of pulling extreme but ill-determined values toward the center of the prior distribution. If the prior has been poorly specified, however, this center may be far from the actual center of the parameter values of interest; estimates of all such parameters will be biased in the same direction. These "ensemble biases" have serious implications for subsequent estimation of examinee individual or population parameters, for while such estimation is resistent to random errors in item parameters, it reflects in direct measure systematic errors in a's and b's, and, through the systematic errors in a's and b's they imply, systematic errors in c's as well.

As a means of overcoming these difficulties, one may introduce the second-stage prior distributions. Experience suggests that item responses of small samples of examinees (less than 2000, say) to short tests (less than 40 items) provide sufficient information to approximate the central tendencies or item parameters through $\underset{\sim}{\mu}_{\xi}$, so that its prior may be diffuse, but not to estimate the convariance matrix $\underset{\sim}{\Sigma}_{\xi}$, so that its prior must be informative. The BILOG computer program (Mislevy & Bock, 1982), for example, fixes $\underset{\sim}{\Sigma}_{\xi}$,

at user-specified values, so that item parameters shrink toward the center of their distribution at a rate controlled by the user, but that center is estimated from the data.

Some of these effects can be illustrated with an analysis of a simulated data set, with responses of 1000 simulated examinees selected at random from a unit normal population to 20 test items. The parameters of the items were also generated from independent normal distributions; for the $\alpha = \log a$, the mean and variance were 0.0 and .5; for $\beta = b$, .5 and 1.0; and for $\gamma =$ logit c, -1.39 and .16. Item parameters were estimated in two ways:

1. Marginal maximum likelihood (MML). Using the BILOG computer computer program, the following likelihood equation was maximized with respect to item parameters $\underset{\sim}{\xi}$ and weights $p_q$ at ten equally spaced quadrature points $X_q$ between -4 and +4:

$$L = \prod_i \sum_q p(\underset{\sim}{u}_i | \underset{\sim}{\alpha}, \underset{\sim}{\beta}, \underset{\sim}{\gamma}, X_q) p_q$$

$$= \prod_i \int_\theta p(\underset{\sim}{u}_i | \underset{\sim}{\alpha}, \underset{\sim}{\beta}, \underset{\sim}{\gamma}, \theta)\, g(\theta)\, d\theta \quad .$$

2. Bayes estimation. To obtain Bayes modal estimates of item parameters, a posterior of similar form was maximized:

$$p(\underset{\sim}{\alpha},\underset{\sim}{\beta},\underset{\sim}{\gamma},\underset{\sim}{p},\underset{\sim}{\mu}_\xi \mid \underset{\sim}{U},\underset{\sim}{\Sigma}_\xi) = \prod_i \sum_q p(\underset{\sim}{u}_i \mid \underset{\sim}{\alpha},\underset{\sim}{\beta},\underset{\sim}{\gamma},X_q) \cdot p_q$$

$$\cdot\ p(\underset{\sim}{\alpha},\underset{\sim}{\beta},\underset{\sim}{\gamma} \mid \underset{\sim}{\mu}_\xi,\underset{\sim}{\Sigma}_\xi) \quad .$$

This is the "floating priors" option of BILOG; the mean vector $\underset{\sim}{\mu}_\xi$ of item parameters is estimated concurrently with the item parameters themselves, but a prior covariance matrix is supplied by the user. BILOG default values of 1.00, 4.00, and 0.25 were employed for $\Sigma_{\alpha\alpha}$, $\Sigma_{\beta\beta}$, and $\Sigma_{\gamma\gamma}$. (These values are intended to be sufficiently mild to affect most parameters minimally when the data supply information about them, but keep all estimates within "reasonable" ranges.) Off-diagonal elements of $\underset{\sim}{\Sigma}_\xi$ were set at zero.

The value of -2 log L under the MML solution was found to be 22,295, while the value obtained by substituting the Bayes estimates into the likelihood function was 22,300. This trivial difference implies that the Bayes estimates explain the observed data nearly as well as the MML estimates.

Indeed, with a few exceptions (more on these below), MML and Bayes estimates of $\underset{\sim}{\alpha}$ and $\underset{\sim}{\beta}$ were quite similar, with $\alpha$'s tending to be shrunken slightly toward their estimated mean of .21.

Estimates of asymptotes were more significantly affected, as seen in Figures 1 and 2. These figures plot generating and estimated values of c, MML and Bayes solutions respectively, against generating values of the quantity b - 2/a, a hueristic index based on the observation that less information is obtained about c as items become easier or less reliable (Lord, 1975). Items with high values of this index are seen to have estimated c's near their generating values under both estimation procedures, but certain items with low values are regressed strongly toward the estimated mean of about .21. To anthropomorphize, we might say that the Bayes solution felt true c's for these items were probably more similar to the c's that it could estimate well than to the atypical and unstable MML values based on sparse information.

---

Insert Figures 1 and 2 about here

---

It is instructive to consider the estimated a's and b's of these items, to see how item parameters can "trade off" against one another. Values for the six items showing the largest differences between MML and Bayes estimated c's are shown in Table 1. The following results may be observed:

Insert Table 1 about here

Item 1 is relatively easy, so that the increased c value obtained by the Bayes solution has little effect on the estimated a and b. As it turns out, the generating c for this item was lower and more atypical than either MML or Bayes obtained, but since most of the examinees were well above the chance level, it did not really matter. Item 4 is similar, in that a large degree of shrinkage of the estimated c on an easy items has little effect on the other parameters. This time (and, the model assumes, more often than not) the Bayes estimate is closer to the true value.

Item 2 shows an extremely high c under MML shrunken back by Bayes procedures to a lower, more nearly correct, value. While the estimated a's are similar, the estimated b under Bayes is correspondingly reduced somewhat, again closer to its true value. The point here is that spuriously over-estimated c's induce spuriously over-estimated b's, a result guarded against in two ways when priors are enforced on both parameters.

Items 3 and 6 show items with high MML a estimates being shrunken back toward their mean under Bayes, and extreme c's correspondingly regressed. Both items are relatively easy, but it is seen that pulling down a spuriously high c (item 3) affects b whereas increasing a spuriously low c (item 6) does not.

Finally, item 5 shows an atypically low c regressing toward its mean, causing a corresponding shift in a away from its mean. The estimated b's are similar under both models.

## Discussion

Maximum likelihood (ML) estimation is justified by its asymptotic properties alone. Taking the data for each parameter at face value no matter how sparse, ML will often yield infinite or implausible parameter estimates in small samples. Thissen and Wainer (1982) suggest that at least for certain parameters, a sample size of 10,000 examinees can be a small sample in the context of the three-parameter logistic IRT model; estimation procedures therefore stand to profit from the incorporation of additional information. The hierarchical Bayesian framework given in the present article supplied such information in a very modest way. In effect, it quantifies beliefs such as

1. if the items for which we can reasonably estimate c's yield values between .1 and .3, then the items for which less information is available probably has c's in this range as well;
2. if most of the items have a's between 1/3 and 3, then the a for this particular item is probably not 957;
3. if all of the other examinees seem to have $\theta$'s between -3 and +3, the $\theta$ for this examinee is probably not $+\infty$, even though he did correctly answer both items he was presented.

Such strictures are implied by the assumption that parameters belong to respective well-behaved populations, the higher-level parameters of which little or nothing need be assumed. The effect of this so-called assumption of exchangeability is to "shrink" estimates from where they would have been under ML toward the centers of the respective populations. (This is always true for unimodal prior distributions, though with multimodal priors certain parameters may be shrunk toward local modes rather than the global mode.)

When it is not reasonable to assume a common population, however, exchangeability is violated. Graphic examples of the absurdities that can result are suggested by proponents as well as critics of "shrunken" estimators. Should one expect to obtain better estimates of the true batting averages of baseball players, for instance, by including data on the price of wheat? The point is that shrinking estimates toward a common center is justified only when a common population best represents the extent of our prior knowledge. The imposition of exchangeability across all units, and estimation procedures that require it, are not strictly appropriate when additional information differentiating the units is at hand.

It is in fact this latter case that typically prevails in educational and psychological measurement. Already known, or available more economically than responses from examinees, is information from several sources:

1. Cognitive processing requirements of items can be specified, at least to some degree. Mental rotation items, for example, can be characterized in terms of the number of degrees the target object has been rotated; differential calculus items, an example discussed by Fischer (1973), can be characterized in terms of the derivations rules they demand for solution.
2. Surface features of items can be identified which can suggest a need for distinguishing subpopulations of items. Free-response and multiple-choice items in the same test may be distinguished, for example, as may be analogy items from vocabulary items in the SAT.
3. Item content can be often be identified. In a test of reading comprehension, one might wish to differentiate items associated with narrative passages, poetry, and documents.
4. Quantitative information, such as percents-correct from pretesting may be available.
5. Examinees may be differentiated with respect to qualitative features such as sex, educational program, or racial/ethnic background; or with respect to quantitative variables such as scores on previously-administered tests.

More comprehensive Bayesian procedures would provide for the utilization of such information. They would also provide for means

of determining when such information makes material differences in item and population parameter estimates.

## References

Andersen, E. B., & Madsen, M. (1977). Estimating the parameters of a latent population distribution. Psychometrika, 42, 357-374.

Ando, A., & Kaufman, O. M. (1965). Bayesian analysis of the independent normal process--neither mean nor precision known. Journal of the American Statistical Association, 60, 347-358.

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. In F. M. Lord & M. R. Novick, Statistical theories of mental test scores. Reading, MA: Addison-Wesley.

Bock, R. D., & Aitken, M. (1981). Marginal maximum likelihood estimation of item parameters: Application of an EM algorithm. Psychometrika, 46, 443-459.

Bock, R. D., & Mislevy, R. J. (1982). Adaptive EAP estimation of ability in a microcomputer environment. Applied Psychological Measurement, 6, 431-444.

Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. Journal of the American Royal Statistical Society, Series B, 39, 1-38.

Efron, B., & Norris, C. (1975). Data analysis using Stein's estimator and its generalizations. Journal of the American Statistical Association, 70, 311-319.

Fisher, G. (1973). The linear logistic test model as an instrument in educational research. Acta Psychologica, 37, 359-374.

James, W., & Stein, C. (1961). Estimation with quadratic loss. Proceedings of the Fourth Berkeley Symposium on Mathematical Probability and Statistics (Vol. 1). Berkeley: University of California Press.

Jeffreys, H. (1961). Theory of probability (3rd ed.) Oxford: Clarendon Press.

Kelley, T. L. (1927). The interpretation of educational measurements. New York: World Press.

Kendall, M. G., & Stuart, A. (1973). The advanced theory of statistics (Vol II., 3rd ed.). New York: Hafner.

Leonard, T. (1982). Comment on 'A simple predictive density function' by Lejeune and Falkenberry. Journal of the American Statistical Association, 77, 657-658.

Lindley, D. V., & Smith, A. F. M. (1972). Bayes estimates for the linear model. Journal of the Royal Statistical Society, Series B, 34, 1-41.

Lord, F. M. (1952). A theory of test scores. Psychometric Monograph, No. 7. Psychometric Society.

Lord, F. M. (1975). Evaluation with artificial data of a procedure for estimating ability and item characteristic curve parameters (RB-75-33). Princeton, NJ: Educational Testing Service.

Lord, F. M. (1980). Applications of item response theory to practical testing problems. Hillsdale, NJ: Erlbaum.

Mislevy, R. J. (1984). Estimating latent distributions. Psychometrika, 49, 359-381.

Mislevy, R. J., & Bock, R. D. (1981, July). Implementation of an EM algorithm in the estimation of item parameters. Paper presented at the IRT/CAT Invitational Conference, Minneapolis, MN.

Mislevy, R. J., & Bock, R. D. (1982). BILOG: Item analysis and test scoring with binary logistic models [Computer program]. Mooresville, IN: Scientific Software.

Novick, M. R., Jackson, P. H., Thayer, D. T., & Cole, N. S. (1972). Estimating multiple regressions in m-groups: A cross-validational study. British Journal of Mathematical and Statistical Psychology, 5, 33-50.

O'Hagan, A. (1976). On posterior joint and marginal modes. Biometrika, 63, 329-333.

Rasch, G. (1960). Probabilistic models for some intelligence and attainment tests. Copenhagen: Danish Institute for Educational Research.

Reiser, M. R. (1981, June). Bayesian estimation of item parameters in the two-parameter logistic model. Paper presented at the annual meeting of the Psychometric Society in Chapel Hill, NC.

Rigdon, S., & Tsutakawa, R. K. (1983). Parameter estimation in latent trait models. Psychometrika, 48, 567-574.

Rubin, D. B. (1980). Using empirical Bayes techniques in the law school validity studies. Journal of the American Statistical Society, 75, 801-827.

Sanathanan, L., & Blumenthal, N. (1978). The logistic model and latent structure. Journal of the American Statistical Association. 73, 794-798.

Stroud, A. H., & Sechrest, D. (1966). Gaussian quadrature formulas. Englewood Cliffs, NJ: Prentice-Hall.

Swaminathan, H., & Gifford, J. A. (1982). Bayesian estimation in the Rasch model. Journal of Educational Statistics, 7, 175-192.

Swaminathan, H., & Gifford, H. A. (1984, in press). Bayesian estimation in the three-parameter logistic model. Psychometrika.

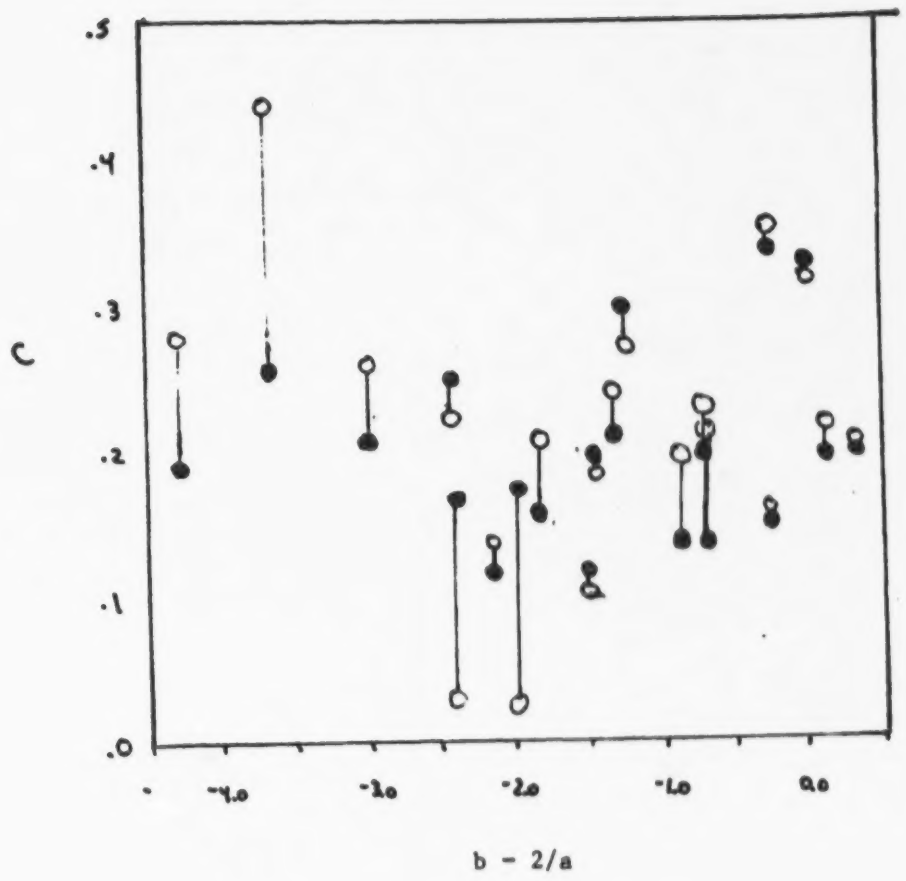
Thissen, D. (1982). Marginal maximum likelihood estimation for the one-parameter logistic model. Psychometrika, 47, 175-186.

Wainer, H., & Thissen, D. (1982). Some standard errors in item response theory. Psychometrika, 47, 397-412.

Table 1

Generating and Estimated Parameters of Selected Items

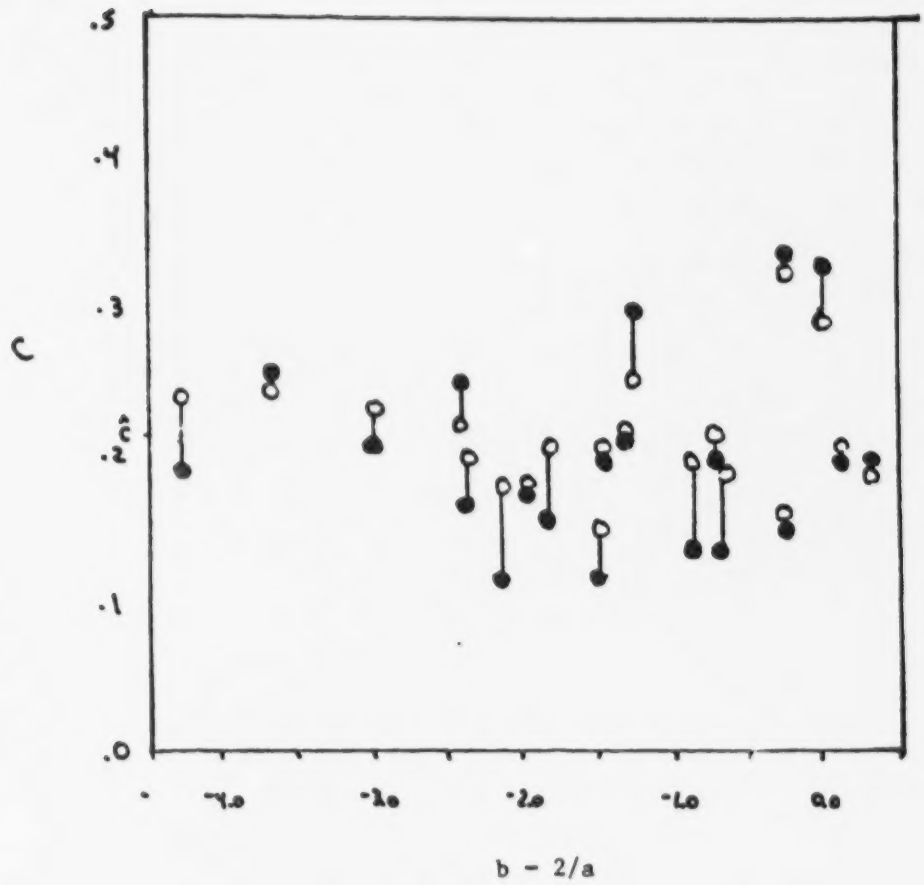
| | a | | | b | | | c | | |
|---|---|---|---|---|---|---|---|---|---|
| Item | True | MML | Bayes | True | MML | Bayes | True | MML | Bayes |
| 1 | 1.1 | 1.2 | 1.3 | -.4 | -.4 | -.3 | .11 | .14 | .17 |
| 2 | .5 | .4 | .4 | .2 | .8 | .6 | .19 | .28 | .24 |
| 3 | .9 | 1.5 | 1.1 | -1.3 | -.6 | -1.0 | .26 | .44 | .27 |
| 4 | 1.4 | 1.2 | 1.4 | -1.0 | -1.2 | -1.0 | .17 | .03 | .19 |
| 5 | 1.5 | 2.2 | 2.4 | -.3 | -.2 | -.2 | .13 | .12 | .14 |
| 6 | 2.5 | 4.5 | 3.4 | -1.1 | -1.2 | -1.1 | .18 | .03 | .18 |




● generating value of c

o estimated value of c

Figure 1. Generating and MML estimated values of c, against generating b - 2/a.




• generating value of c

o estimated value of c

Figure 2. Generating and Bayes estimated values of c, against generating b - 2/a.